Smart Intelligent Computing and Communication Technology
V.D. Ambeth Kumar et al. (Eds.)

doi:10.3233/APC210010

Design and Analysis of Hybrid Classification Approaches Across Multi-Relational Decision Tree Learning Algorithm

Tressa Michael[a,1],
[a,1]*Assistant Professor, Dept of ECE, Rajagiri School of Engineering and Technology, Kerala, India*

Abstract. In today's day of Modern era when the data handling objectives are getting bigger and bigger with respect to volume, learning and inferring knowledge from complex data becomes the utmost problem. Almost all of the real- world information are maintained under a relational fashion holding multiple relations unlike orthodox approaches containing single relational as a whole. Moreover several fields viz. biological informatics, microbiology, chemical computations needed some more dependable and expressive approach which can provide more sophisticated results with faster speed. Hence in context with multi-relational data mining in which data is directly retrieved from different records without dumping into single table, we have described a novel approach of improved Multi-Relational Decision Tree Learning Algorithm based on the implementations. In this paper provided a comparative study of the aforementioned approach in which we have taken certain results from the literature review. Experiments mainly includes results from widely used datasets viz. Mutagenesis database which demonstrates that Multi-Relational Decision Tree Learning Algorithm provides a promising alternative from previous conventional approaches such as Progol, FOIL, and Tilde.

Keyword. Data Mining, KDO, Learning Model, MRDTL, Relational Attribute, Relational Table, UML.

1. Introduction

The word 'Data Mining' implies to a process of retrieving of important and useful insights through huge large detailed collections of data. Data mining can also be understood as the process of finding new and meaningful patterns through huge amounts of raw and unsettled data. These data amounts could be either reserved or gathered into the respective databases or warehouses another repository sources [1]. Alternatively the action of retrieving unidentified, valid, actionable meaningful data from big data chunks and then utilizing that information to make strategic decisions may also be termed as Data mining. It involves sorting through large amount of data and picking up relevant information and useful patterns in data using different

[1]Tressa Michael, Assistant Professor, Dept of ECE, Rajagiri School of Engineering and Technology, Kerala, India, Email:tressa_m@rajagiritech.edu.in

algorithms to uncover and derive hidden facts. It identifies local obvious patterns within data that go beyond the scope of simple analysis with the help appropriate and sophisticated algorithms [2]. Data mining encompasses various ways taken through different disciplines viz. business intelligence, financial analysis, machine learning, computational and high-performance computing, recognising patterns, visualization of data visualization, retrieval of meaningful information etc. As huge volume of information is increasing exponentially every year, the data acquired is also getting bigger and hence its data mining is turning to be a valuable tool in order to convert the data into meaningful information. This continuous act of research and development has changed mining of data pattern into its refined form.

1.1 Functionalities of Data Mining

The major fundamental functionalities of data mining can be describes as:
(i)Prediction or Predictive Data Mining: It perform inferences on the current data to make prediction and utilizes previously existing variables in a way to decide the values which satisfies the interest.(ii) Description: It characterizes the properties of information through the database which aims on searching patterns which describes the data and then consequent presentation is developed for user interpretation.

The very major and indispensable ingredient for any Data Mining is the database. Any database represents a managed, properly organized and basically huge pile of detailed but raw information regarding some stream or topic linked to outer world. The vital and main work of data mining involves the auditing of the database for pattern consistency which can provide a more sober inference of the topic proposed by the database. In this process we basically take that it includes of a group or pair of individuals [3]. Relying on the topic of the domain, the individual entities may be anonymous, from an account holder to chemical compounds.

2. Literature Review

Ganganwar (2012) represents a brief review of existing techniques for class imbalance that are data resampling approaches and algorithmic approaches. In this data level approaches rebalance the classes artificially by either oversampling or under sampling. In the algorithmic approaches instead of rebalancing, algorithms of classifiers are changed such that making classifiers biased towards the class of interest that is the minority class. Thus by algorithmic approaches more minority instances are classifying but it is very tough to build.

Lopez et al. (2013) carry out a thorough discussion on the data intrinsic problems which lead to poor performance of classifier. Data intrinsic problems includes small disjunctive problems, overlapping between the classes, density of classes in the training dataset, noisy data, borderline instances and data shift between distribution of training and test dataset. It will also introduce several approaches and recommendations for solving this intrinsic problem with the class imbalance problems. And it will also show some experimental examples of the dataset which includes this data intrinsic problems. It would be obvious that successful extraction of data will require a larger database [4]. This article would include treatment for issues surrounding the objective structure of objects. Salters-Wise database provides detailed information about the molecules and tables for the molecules and uses Mendeleev's table of the periodic elements as context

details for the periodic elements (Mendeleev tables) [5] . It is important to learn from and obtain a foothold while Lerner is being rejected (blokila, 1998). Multiple Example Learning Exercises is one where each teaching example is represented by multiple instances (or feature vectors) and they can be applied to all molecules for which molecules have undefined numbers of properties. Many ILP methods [6] have been applied to deal with different issues and will continue to be used in the future.[19-22]

In each of the cases listed above, a range of techniques suggested still rely on their proposed components for the extraction of relevant data. For e.g., ILP Cloudier, TILD and ICL are first order upgrade for proposed data- mining algorithms to enforce a Reinforcement Learning and Reinforcement Learning system by passing rules on affiliation, grouping, and decision-making bodies [7]. The related models for relation domains in the Bayesian networks have eventually been generalized to include the TILDE logical decision-making tree, which includes the decision-making tree introduced in[8], in line with the proposed decision tree algorithm.[16-18] Many adopt a method of updating learning algorithms to conform to the relevant learning, as stated in [9].

3. Methodology

3.1 Relational Data Learning

The research in Knowledge Discovery in Databases has been primarily directed to attribute-value learning in which one is described through a fixed set (tuple) given with their values. Database or dataset is seen in the form of table (relation) in which every row corresponds to an instance and column represents an attribute respectively. Base language used lies or based on propositional logic, which can show in the form "Attribute ⊕ Value" where ⊕ can be from a fixed set of predefined operators viz. {<,>, ≤, ≥, =}.

3.2 Proposed Approaches for Relational Data Mining

Several techniques proposed for relational data mining are discussed below: Inductive Logic Programming. The ILP's principle paradigm comprises the usage of Induction and Logic Programming. Prolog is as the principle representative language for implementation and deductive reasoning [**10**].The whole process of reducing and deduction basically depends at usage of significant inferring rules. Its presumption includes rules depending statistical support data. Proposed Unified Modelling Language (UML) which can be used as single language for specifying different ILP engines. The main work or aim concentrates to change the whole logic which is depending on formalism which specify the language as proper and sophisticated language which is simple to comprehend and implement consequently, and can be implemented by a different ILP systems, and can also be able to depict the problem's complexity with simple and in clear way. Through this, even non-commissioned users can frame query and can employ different ILP engines, which best fits their requirements. Another problem regarding usage of ILP learning is its confined capabilities for the relational database. Usage of relational engines elevates the ILP system's working strengths. In [11,12], three methods for connecting ILP and relational databases were proposed are explained.

4. Experiments and Results

With all the approaches proposed previous to MRDTL, there are numerous results in consideration with multi-relational data mining which apply on several ILP engines in order to infer information from multiple database. We will use a widely used database i.e. Mutagenesis Database [13], which is consistently used in several ILP research and will provide a comparative graphical study with the results obtained from the literature.

4.1 Comparison between MRDTL and previous approaches

The Mutagenesis database holds of 230 entities of molecules which are further sub classified into two groups with 188 molecules which acts as the entities which are regression friendly and 42 molecules acting as regression unfriendly. In friendly section it contains 4893 entities which represents atoms schema and 5243 entities which represents bonds and in other set it contains 1001 and 1066 respectively [14]. The mutagenesis database consists of five levels of background knowledge viz. B0 to B4. A graphical comparison of different previous approaches viz. Progol, FOIL, TILDE with MRDTL [15] using these background knowledge is performed in context with accuracy and execution time. With the results taken from the literature, we had prepared Table 1 comparing the results on the accuracy factor of MRDTL with previous approaches.

Table 1. Accuracy Comparisons with Mutagenesis Database

	Accuracy %)			
System	**B0**	**B1**	**B3**	**B4**
Progol	79	86	86	88
Foil	62	62	83	82
Tilde	75	78	85	86
MRDTL	67	89	86	89

Due to the dissimilarity in hardware and software decencies employed in the different streams it is almost impossible to implement these algorithms in real world. Despite the fact that certain ILP depending engines viz. Tilde and FOIL are present online through the internet, but all of them works only on the Solaris OS versions only. Hence, we are not able to could reproduce those facts in our working paradigms. So using the values provided in the literature we had provided a graphical comparison of MRDTL with previous approaches. With the results obtained, it is evident from the Fig. 1, that MRDTL approach is much more consistent, accurate and signifies its better approach.

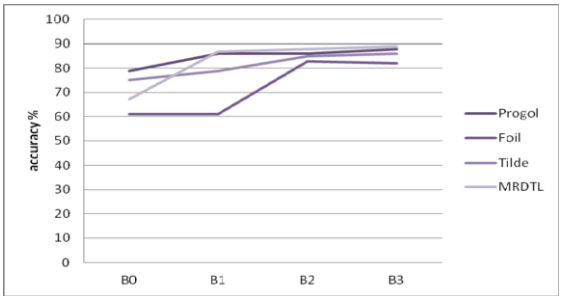

Figure 1. Graphical Comparison (accuracy) of MRDTL with previous approaches

We had prepared a Table 1 using the results comparing the running time factor of MRDTL with previous approaches.

Table 2. Comparison on Performance Accuracy

Dataset	MRDTL accuracy (%)	Hybrid MRDTL accuracy (%)
Mut dataset	87.5	86
Loc dataset	76.11	72.10
Function dataset	91.44	93.6
Thrombosis dataset	98.1	99.28

These results are graphically mapped, it is evident from the graph that Hybrid Multi-Relational Decision Tree Learning Algorithm more accurate than its predecessor.

5. Conclusion

In this paper we have provided a comparative study of the aforementioned approach in which we have taken certain results from the literature. Our experiments mainly includes results from widely used datasets viz. Mutagenesis database which demonstrates that MRDTL algorithm provides a promising alternative from previous conventional approaches such as Progol, FOIL, and Tilde. Our research also points out certain major drawbacks of this approach which causes hindrance in its smooth working. Consequently we introduced a much more sophisticated and deserving approach i.e. improved Multi-Relational Decision Tree Learning Algorithm which overcomes with those drawbacks and other anomalies. Slow performance: improved Multi-Relational Decision Tree Learning Algorithm provides certain methods which reduces its execution time. Experimental results on different datasets provide a clear indication that improved Multi-Relational Decision Tree Learning Algorithm is comprehensively a better approach.

References

[1] Kersting, K. and De Raedt, L. Bayesian Logic Programs. Proceedings of the Work-in Progress Track at the 10th International Conference on Inductive Logic Programming (2000).
[2] Pfeffer, A. A Bayesian Language for Cumulative Learning, Proceedings of AAAI 2000. Workshop on

Learning Statistical Models from Relational Data, AAAI Press (2000).
[3] Krogel, M. and Wrobel, S. Transformation-Based Learning Using Multirelational Aggregation. Proceedings of the 11th International Conference on Inductive Logic Programming, vol. 2157(2001).
[4] Getoor, L., Multi-relational data mining using probabilistic relational models: research summary. Proceedings of the First Workshop in Multi-relational Data Mining (2001).
[5] Knobbe, A. J., Siebes, A., Blockeel, H., and Vander Wallen, D .Multi-relational data mining using UML for ILP .Proceedings of the First Workshop in Multirelational Data Mining, 2001.
[6] Hector Ariel Leiva . A Multi-Relational Decision Tree Learning Algorithm .M.S. thesis, Department of Computer Science. Iowa State University (2002).
[7] Y. Sun, M.S. Kamel, Y. Wang .Boosting for Learning Multiple Classes with Imbalanced Class Distribution .Proc. Int'l Conf. Data Mining, pp. 592-602. 2006.
[8] N. Abe, B. Zadrozny, J. Langford. An Iterative Method for Multi-Class Cost-Sensitive Learning. Proc.ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 3-11, 2004
[9] K. Chen, B.L. Lu, J. Kwok. Efficient Classification of Multi-Label and Imbalanced Data Using Min-Max Modular Classifiers. Proc. World Congress on Computation Intelligence—Int'l Joint Conf. Neural Networks, pp. 1770-1775, 2006.
[10] Mr. Ankit Kumar, Dr. Dinesh Goyal, Mr. Pankaj Dadheech, (2018). A Novel Framework for Performance Optimization of Routing Protocol in VANET Network.Journal of Advanced Research in Dynamical & Control Systems, Vol. 10, 02-Special Issue, 2018, pp-2110-2121, ISSN: 1943-023X.
[11] Mr. Pankaj Dadheech, Dr. Dinesh Goyal, Dr. Sumit Srivastava, Mr. Ankit Kumar, (2018). A Scalable Data Processing Using Hadoop & MapReduce for Big Data. Journal of Advanced Research in Dynamical & Control Systems, Vol. 10, 02-Special Issue, 2018, pp-2099-2109, ISSN: 1943-023X.
[12] Pankaj Dadheech, Dinesh Goyal, Sumit Srivastava & C. M. Choudhary, (2018). An Efficient Approach for Big Data Processing Using Spatial Boolean Queries. Journal of Statistics and Management Systems (JSMS), 21:4, 583-591.
[13] A. Kumar and M. Sinha (2014).Overview on vehicular ad hoc network and its security issues. International Conference on Computing for Sustainable Global Development (INDIACom), pp. 792-797. doi: 10.1109/IndiaCom.2014.6828071.
[14] Pankaj Dadheech, Ankit Kumar, Chothmal Choudhary, Mahender Kumar Beniwal, Sanwta Ram Dogiwal & Basant Agarwal (2019). An Enhanced 4-Way Technique Using Cookies for Robust Authentication Process in Wireless Network. Journal of Statistics and Management Systems, 22:4, 773-782, DOI: 10.1080/09720510.2019.1609557.
[15] Ankit Kumar, Pankaj Dadheech, Vijander Singh, Linesh Raja & Ramesh C. Poonia (2019).An Enhanced Quantum Key Distribution Protocol for Security Authentication. Journal of Discrete Mathematical Sciences and Cryptography. 22:4, 499-507. DOI: 10.1080/09720529.2019.1637154.
[16] VDAK, S. Sharmila, Abhishek Kumar, A. K. Bashir, Mamoon Rashid, Sachin Kumar Gupta & Waleed S. Alnumay .A novel solution for finding postpartum haemorrhage using fuzzy neural techniques. Neural Computing and Applications (2021) (https://doi.org/10.1007/s00521-020-05683-z)
[17] AnkitKumar,VijayakumarVaradarajan,AbhishekKumar, PankajDadheech, Surendra SinghChoudhary, V.D. AmbethKumar, B.K.Panigrahi, Kalyana C.Veluvolug. Black hole attack detection in vehicular ad-hoc network using secure AODV routing algorithm. Microprocessors and Microsystems, In Press,(https://doi.org/10.1016/j.micpro.2020.103352)
[18] Ambeth Kumar.A Cognitive Model for Adopting ITIL Framework to Improve IT Services in Indian IT Industries. Journal of Intelligent Fuzzy Systems. (DOI: 10.3233/JIFS-189131)(Accepted - Inpress)
[19] V.D.AKumar .Efficient Data Transfer in Edge Envisioned Environment using Artificial Intelligence based Edge Node Algorithm. Transactions on Emerging Telecommunications Technologies (Accepted - Inpress)(DOI: 10.1002/ett.4110)
[20] V. D. AK ,S. Malathi,Abhishek Kumar,Prakash M and Kalyana C. Veluvolu .Active Volume Control in Smart Phones Based on User Activity and Ambient Noise .Sensors 2020, 20(15), 4117; https://doi.org/10.3390/s20154117
[21] AK S. Malathi R. Venkatesan K Ramalakshmi, Weiping Ding, Abhishek Kumar .Exploration of an innovative geometric parameter based on performance enhancement for foot print recognition. Journal of Intelligent and Fuzzy System . vol. 38, no. 2, pp. 2181-2196. 2020. DOI: 10.3233/JIFS-190982
[22] B. Aravindh; V.D.Ambeth Kumar; G. Harish; V. Siddartth, " A novel graphical authentication system for secure banking systems", IEEE (ICSTM), Pages: 177 – 183, 2-4 Aug. 2017, DOI: 10.1109/ICSTM.2017.8089147